Web Spam, Propaganda and Trust

Panagiotis T. Metaxas
Computer Science Department
Wellesley College
Wellesley, MA 02481, USA
pmetaxas@wellesley.edu

Joseph DeStefano
Math and Computer Science Department
College of the Holy Cross
Worcester, MA 01610, USA
joed@mathcs.holycross.edu

ABSTRACT

Web spamming, the practice of introducing artificial text and links into web pages to affect the results of searches, has been recognized as a major problem for search engines. It is also a serious problem for users because they are not aware of it and they tend to confuse trusting the search engine with trusting the results of a search.

In this paper, we first analyze the influence that web spam has on the evolution of the search engines and we identify the strong relationship of spamming methods to propagandistic techniques in society. Our analysis provides a foundation to understanding why spamming works and offers new insight on how to address it. In particular, it suggest that one could use anti-propagandistic techniques in the web to recognize spam. The second part of the paper demonstrates such a technique, called backwards propagation of distrust.

In society, recognition of an untrustworthy message (in the opinion of a particular person or other social entity) is a reason for questioning the entities that recommend the message. Entities that are found to strongly support untrustworthy messages become untrustworthy themselves. So, social distrust is propagated backwards for a number of steps. Our algorithm simulates this social behavior on the web graph.

In our algorithm, starting from an untrustworthy (according to the end user) site s, we examine its *trust neighborhood*, that is, the neighborhood of sites that link to s in a few steps. Evaluating the sites-members of the neighborhood we identify a biconnected component (BCCs) with a high percentage of untrustworthy sites. BCCs are formed when there are multiple directed paths to reach s, thus indicating a concerted effort to promote s. This is not the case when starting from a trustworthy site.

Our tool explores thousands of nodes within minutes and could be deployed at the browser-level, making it possible to resolve the moral question of who should be making the decision of weeding out spammers in favor of the end user.

Our approach can lead to browser-level web spam filters that work in synergy with the powerful search engines to deliver personalized, trusted web results.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.m [**Information Storage and Retrieval**]: Miscellaneous

General Terms

Algorithms, Experimentation, Social Networks, Propaganda, Trust

Keywords

search, Web graph, link structure, PageRank, HITS, Web spam

1. INTRODUCTION

Web spamming is often defined as the practice of manipulating web pages in order to cause search engines to rank some web pages higher than they would without any manipulation[1]. Spammers aim at search engines, but target the end users. Their motive is usually commercial, but can also be political, or religious.

One of the reasons behind the users' difficulty to distinguish trustworthy from untrustworthy information comes from the success that both search engines and spammers have enjoyed in the last decade. Users have come to trust search engines as a means of finding information, and spammers have successfully managed to get them to transfer that trust to the results of each search.

From their side, the search engines have put considerable effort in delivering spam-free query results and have developed sophisticated ranking strategies. Two such ranking strategies that have received major attention are the well-known PageRank [6] and HITS [28] algorithms. Achieving high PageRank has become a sort of obsession for many companies' IT departments, and the *raison d'être* of spamming companies. Some estimates indicate that at least 8% of all pages indexed is spam [12] while experts consider web spamming the single most difficult challenge web searching is facing today. [23]. Search engines typically see web spam as an interference to their operations and would like to restrict it, but there can be no algorithm that can recognize spamming sites based solely on graph isomorphism [5].

First, however, we need to understand why spamming works beyond the technical details, because spamming is a social problem first, then a technical one. In this paper we show its extensive relationship to social propaganda, and evidence of its influence on the evolution of search engines.

[1]We should mention here that there is not a complete agreement on the definition of web spam among authors, which leads to some confusion. Moreover, to people unfamiliar with web spam, the term is mistaken for email spam. A more descriptive name for it would be "search engine ranking manipulation."

Our approach can explain the reasons web spamming has been so successful and suggest new algorithmic ways of dealing with it. Finally, we discuss what we believe should be a frame for the long-term approach to web spam.

The rest of this paper is organized as follows. The next section gives an overview of the problem of web spamming and information reliability for a general audience. Section 3 discussed the relationship between webgraph and the trust social network while the following section analyzes the evolution of search engines as their response to spam. Section 5 describes the backward propagation of distrust method and the following section presents some of our experimental results running this algorithm. Section 7 discusses some related research and the final section has our conclusions and some discussion of future directions of this work.

2. BACKGROUND

The web has changed the way we inform and get informed. Every organization has a web site and people are increasingly comfortable accessing it for information for any question they may have. The exploding size of the web necessitated the development of search engines and web directories. Most people with online access use a search engine to get informed and make decisions that may have medical, financial, cultural, political, security or other important implications [10, 39, 24, 32]. Moreover, 85% of the time, people do not look past the first ten results returned by the search engine [38]. Given this, it is not surprising that anyone with a web presence struggles for a place in the top ten positions of relevant web search results. The importance of the top-10 placement has given birth to a new industry, which claims to sell know-how for prominent placement in search results and includes companies, publications, and even conferences. Some of them are willing to bend the truth in order to fool the search engines and their customers, by creating web pages containing web spam [12].

The creators of web spam are often specialized companies selling their expertise as a service, but can also be the web masters of the companies and organizations that would be their customers. Spammers attack search engines through text and link manipulations [23, 19]:

- **Text spam**: This includes excessively repeating text and/or adding irrelevant text on the page that will cause incorrect calculation of page relevance; adding misleading meta-keywords or irrelevant "anchor text" that will cause incorrect application of rank heuristics.

- **Link spam**: This technique aims to change the perceived structure of the webgraph in order to cause incorrect calculation of page reputation. Such examples are the so-called "link-farms", "mutual admiration societies", page "awards", domain flooding (plethora of domains that re-direct to a target site), etc.

Both kinds of spam aim to boost the ranking of spammed web pages. Sometimes **cloaking** is included as a third spamming technique [23, 20]. Cloaking aims to serve different pages to search engine robots and to web browsers (users). These pages could be created statically or dynamically. Static pages, for example, may employ hidden links and/or hidden text with colors or small font sizes noticeable by a crawler but not by a human. Dynamic pages might change content on the fly depending on the visitor, submit millions of pages to "add-URL" forms of search engines, etc. We consider the false links and text themselves to be the spam, while, strictly speaking, cloaking is not spam, but a tool that helps spammers hide their attacks.

Since anyone can be an author on the web, these practices have naturally created a question of *information reliability*. An audience used to trusting the written word of newspapers and books is unable, unprepared or unwilling to think critically about the information obtained from the web. A recent study [17] found that while college students regard the web as a primary source of information, many do not check more than a single source, and have trouble recognizing trustworthy sources online. In particular, two out of three students are consistently unable to differentiate between facts and advertising claims, even "infomercials." Very few of them would double-check for validity. At the same time, they have considerable confidence in their abilities to distinguish trustworthy sites from non-trustworthy ones, especially when they feel technically competent. We have no reason to believe that the general public will perform any better than well-educated students. In fact, a recent analysis of internet related fraud by a major Wall Street law firm [10] puts the blame squarely on the investors for the success of stock fraud cases.

3. THE WEBGRAPH AS A SOCIAL NET

The web is typically represented by a directed graph [8]. The nodes in the webgraph are the pages (or sites) that reside on servers on the internet. Arcs correspond to hyperlinks that appear on web pages (or sites). *Web spammers are trying to alter the web graph in ways beneficial to them.*

The theory of social networks of Trust [40] also uses directed graphs to represent relationships between social entities. The nodes correspond to social entities (people, institutions, ideas). Arcs correspond to recommendations between the entities they connect. *Propagandists are trying to alter the trust social net in ways beneficial to them.*

This connection is more than just a similarity in descriptions. The web itself is a social creation, and both PageRank and HITS are socially inspired ranking algorithms. [6, 28, 36]. Socially inspired systems are subject to socially inspired attacks, however. Not surprisingly then, the theory of propaganda detection [31] can provide intuition into the dynamics of the web graph. First developed in the beginning of World War II by the Institute for Propaganda Analysis [15, 31], the theory of propaganda detection identifies several techniques that propagandists often employ in order to manipulate perception. Name calling, glittering generalities, testimonial, bandwagon and transfer are the more well-known of them.

PageRank is based on the assumption that the reputation of an entity (a web page in this case) can be measured as a function of both the number and reputation of other entities linking to it. A link to a web page is counted as a "vote of confidence" to this web site, and in turn, the reputation of a page is divided among those it is recommending [2]. The implicit assumption is that hyperlink "voting" is taking place independently, without prior agreement or central

[2]Since HTML does not provide for "positive" and "negative" links, all links are taken as positive. This is not always true, but is considered a reasonable assumption. Recently, Google introduced the "nofollow" attribute for hyperlinks, but it is very unlikely that web spammers will use it.

Graph Theory	*Web Graph*	*Trust Social Network*
node	web page or site	social entity
node weight	rank (accord. to SE)	reputation (accord. to user)
node weight computation	ranking formula	based on top recommenders
	automatic	on demand
arc	hyperlink	trust opinion
arc meaning	vote of confidence	recommendation
arc weight	degree of confidence	degree of entrustment
arc weight computation	ranking formula	arbitrary, semi-consistent
arc weight range	$[0 \dots 1]$	$[distrust \dots trust]$

Table 1: Graph theoretic correspondence between the Webgraph and the Trust Social Network.

control. Spammers, like social propagandists, form structures that are able to gather a large number of such "votes of confidence" by design, thus breaking the assumption of independence in a hyperlink.

Table 1 has the correspondence between graph theoretic terms, the web graph according to a search engine, and the trust social network of a particular user.

4. EVOLUTION OF SEARCH ENGINES

In the early 90's, when the web numbered just a few million servers, the **first generation** search engines were ranking search results using classic information retrieval techniques: the more rare words two documents share, the more similar they are considered to be. [37, 22] A search query Q is simply a short document and the results of a search for Q are ranked according to their (normalized) similarity to the query.

The first attack to this "$tf.idf$ ranking," as it is known, came from within the search engines. Around 1995, search engines started selling search keywords to advertisers as a way of generating revenue: If a search query contained a "sold" keyword, the results would include targeted advertisement and a higher ranking for the link to the sponsor's web site. This is the first time we have a socially inspired ranking, which follows marketing practices of the real world.

Mixing search results with paid advertisement raised serious ethical questions, but also showed the way to financial profits to spammers who started their own attacks by creating pages containing many rare keywords to obtain a higher ranking score. In terms of propaganda theory, the spammers employed a variation of the technique of *glittering generalities* to confuse the first generation search engines [31, pg. 47]:

The propagandist associates one or more suggestive words without evidence to alter the conceived value of a person or idea.

To avoid spammers search engines would keep secret their exact ranking algorithm. Secrecy is no defense, however, since secret rules were figured out by experimentation and reverse engineering. (e.g., [35, 33]).

Second generation search engines started employing more sophisticated ranking techniques in an effort to nullify the effects of glittering generalities. One of the more successful techniques was based on the "link voting principle": Each web site s has value equal to its "popularity", which is influenced by the set B_s of sites pointing to site s. Lycos became the champion of this ranking technique and had its own popularity skyrocket around 1996.[34]. Doing so, it was also distancing itself from the ethical questions introduced by combining advertising with ranking.

Unfortunately, this ranking method did not succeed in stopping spammers either. Spammers started creating clusters of interconnected web sites that had identical or similar contents with the site they were promoting, which subsequently became known as "link farms" (LF). The link voting principle was socially inspired, so spammers used the well known propagandistic method of *bandwagon* to circumvent it [31, pg. 105]:

With it, the propagandist attempts to convince us that all members of a group to which we belong are accepting his program and that we must therefore follow our crowd and "jump on the band wagon".

Similarly, the spammer is promoting the impression of a high degree of popularity by inter-linking many internally controlled sites that will eventually all share high ranking.

The introduction of PageRank in 1998 was a major development for search engines, because it seemed to provide a more sophisticated anti-spamming solution. Under PageRank, not every link contributes equally to the "reputation" of a page. Instead, links from highly reputable pages contribute much higher than links from other sites. That way, the site networks developed by spammers would not influence much their PageRank, and Google became the search engine of choice. HITS is another socially-inspired ranking which has also received a lot of attention. [28]. The HITS algorithm divides the sites related to a query between "hubs" and "authorities". Hubs are sites that contain many links to authorities, while authorities are sites pointed to by the hubs and they both gain reputation.

PageRank and HITS marked the development of the **third generation** search engines [3]. Unfortunately, spammers have again found ways of circumventing them. In PageRank, a page enjoys absolute reputation: its reputation is not restricted on some particular issue. Spammers deploy sites with expertise on irrelevant subjects, and they justifiably acquire high ranking on their expert sites. Then they bandwagon their networked sites with the expert sites, creating a "mutual admiration society" (MAS). This is the well-known propagandistic technique of *testimonials* [31, pg. 74]:

Well known people (entertainers, public figures, etc.) offer their opinion on issues about which they are not experts.

HITS has also shown to be highly spammable by this tech-

[3][7] considers the search engines in our 2nd and 3rd generation to be in the same group. We believe that both the ranking and attack methods put them in different categories.

nique due to the fact that its effectiveness depends on the accuracy of the initial neighborhood calculation.

The table below summarizes our findings for the first three generations of search engines and the correspondence between web spam and social propaganda.

SE	Ranking	Spamming	Propaganda
1st Gen	Doc Similarity	keyword stuffing	glittering generalities
2nd Gen	+ Site popularity	+ link farms	+ bandwagon
3rd Gen	+ Page reputation	+ mutual admiration societies	+ testimonials

Web search corporations are reportedly busy developing the engines of the next generation [7]. The new search engines hope to be able to recognize "the need behind the query" of the user. Given the success the spammers have enjoyed so far, one wonders how will they spam the fourth generation engines. Is it possible to create a ranking that is not spammable? Put another way, can the web as a social space be free of propaganda? Seen in this light, it appears that we are trying to create in cyberspace what societies have not succeeded in creating in their social space. This may not be possible. However, we can learn to live in a web with spam as we live in society with propaganda, given appropriate education and technology.

5. AN ANTI-PROPAGANDISTIC METHOD

Web Spam seems to be the driving force behind the evolution of search engines in their effort to provide quality results. So far, the battle with web spam is only waged at the search engine level, though the end users are the ones affected directly by it. When users query a popular search engine for questions that happen to be the target of unreliable advertisement (e.g., "Can human growth hormone increase muscle mass?") or happen to be controversial in nature (e.g., "is ADHD a real disease?"), they find plethora of responses that can be considered untrustworthy. For example, the first query provides almost exclusively links to human growth hormone (hGH) products that, among other benefits, would significantly increase muscle mass without increased exercise, decrease fat without change in diet or habits, enhance sexual performance, increase the good cholesterol while decreasing the bad, re-grow hair, decrease blood pressure, remove wrinkles, and increase memory retention. Similarly, in the second query one finds an unbalanced view of attention-deficit, hyperactivity disorder (ADHD) that does not include the opinion of major institutions such as the American Psychiatric Association or clinicians in major research universities. To the inexperienced user it may appear that the search engine promotes untrustworthy, unreliable or unbalanced views. What really happens, of course, is that these queries have been the target of spammers.

Since spammers employ propagandistic techniques, it makes sense to design anti-propagandistic methods for defending against them. These methods need to be user-initiated. We are considering trustworthiness to be a personal decision, not an absolute quality of a site. One person's gospel is another's political propaganda, and our goal is to design

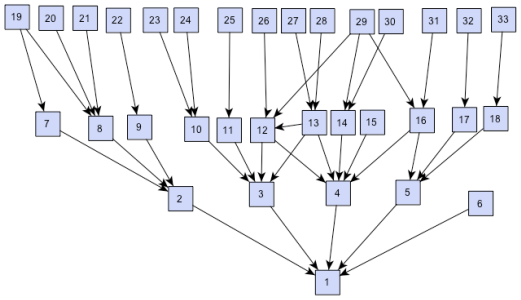

Figure 1: An example of a breadth-first search tree in the trust neighborhood of site 1. Note that some nodes have multiple paths to site 1.

methods that help individuals make more informed decisions about the quality of the information they find on the web.

Here is one way that people defend against propaganda in every day life:

In society, when an untrustworthy recommendation is detected, it gives us a reason to reconsider the trustworthiness of the recommender. Recommenders who strongly support an untrustworthy recommendation become untrustworthy themselves.

This process is selectively repeated a few times, propagating the distrust backwards to those who strongly support the recommendation. The results of this process become part of our belief system and are used to filter future information.

We set out to test whether a similar process might work on the web. Our algorithm takes as input the URL of the server s containing a page that the user determined to be untrustworthy. This page could have come to the user through web search results (like the ones above) or via the suggestion of some trusted associate (e.g., a society that the user belongs to).

Starting from s we build a breadth-first search (bfs) tree of the sites that link to s in a few "clicks" (Figure 1). We do not explore the web neighborhood directly in this step. Instead, we use the Google API [16] for finding the backlinks. We call the directed graph that is revealed by the backlinks, the "trust neighborhood" of s.

The question arizes on whether we should distrust all of the sites in the trust neighborhood of s or not. Is it reasonable to become suspicious of almost every site pointing to s? They are "voting in confidence" after all. Such a radical approach is not what we do in everyday life. Rather, we selectively propagate distrust only to those that most strongly support an untrustworthy recommendation. Thus, we decided to take a conservative approach and examine only those sites that show a more concerted effort in supporting s. In particular, we focused on the biconnected component (BCC) that includes s (Figure 2).

A BCC is a graph that cannot be broken into disconnected pieces by deleting any single vertex. An important characteristic of the BCC is there are at least two independent paths from any of its vertices to s. (Strictly speaking, the BCC is computed on the undirected graph. But since the trust neighborhood is generated through the bfs, every node has directed paths to s.) We view the existence of multiple paths as evidence of strong support of s.

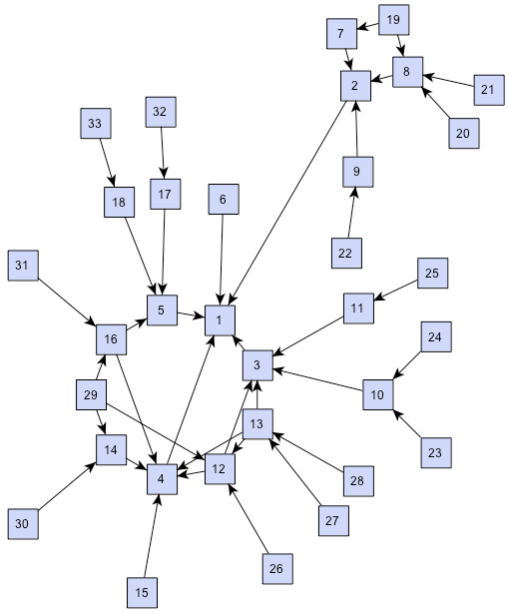

Figure 2: The BCC of the trust neighborhood of site 1 is drawn in a circular fashion for clarity. Nodes 3, 13, 12, 4, 14, 29, 16 and 5, have multiple paths to s.

More formally, the algorithm is as follows:

```
Input: Untrustworthy site s.

S = {s}
Using BFS for depth D do:
   Find the set U of sites linking to sites in S
      using the Google API (up to B backlinks / site)
   Ignore blogs, directories, edu's
   S = S + U

Compute and output the BCC of S that includes s
```

To be able to implement the above algorithm at the browser side, we restrict the following parameters: First, the BFS's depth D is set to 3. We are not interested in exploring a large chunk of the web, just a small neighborhood around s. Second, we limit the number B of backlink requests from the Google API to 30 per site. Finally, we introduced in advance a set of *stop sites* that are not to be explored further. A stop site is one that should not be included in the trust neighborhood either because the trustworthiness of such a site is irrelevant, or because it cannot be defined. In the first category we placed URLs of educational institutions (domains ending in .edu). Academicians are not in the business of pointing to commercial sites. When they do, they do not often convey trust in the site. In the latter we placed a few well known Directories (URLs ending in yahoo.com, dmoz.org, etc.) and Blog sites (URLs containing the string 'blog' or 'forum'). Anyone can put an entry into an unsupervised blog or directory. No effort to create an exhaustive list of blogs or directories was made.

With these restrictions, our algorithm can be implemented on an average workstation and produce graphs with up to a few thousand nodes within minutes. Note that the slowest step is the query of the backlinks. More recently, a threaded version of the program can explore several thousand sites in minutes.

6. EXPERIMENTAL RESULTS

In our experiments, we examined the trust graphs of six untrustworthy and two trustworthy sites, collected from the search results of the first hGH query. In the table 2 below these sites are labeled as U-1 to U-6 and T-1 to T-2, respectively. See Figure 3 for an example of one such site (U-1). We run the experiments between September 17 and November 5, 2004. We should note here that all sites have comparable PageRank. In fact, all but U-1 and T-1 have PageRank 5. The remaining two sites have PageRank 6.

To determine the trustworthiness of each site we had an evaluator look at a sample of the sites of the BCC. Due to the effort involved, only a randomly chosen 20% of the total 1,164 BCC sites were evaluated. (Optimally, we would like to evaluate the whole neighborhood, and we expect that this would strengthen our results, but this would require increasing the manual work by an order of magnitude.) A site was then classified as either Trustworthy, Untrustworthy, or Non-determined. The last category includes a variety of sites for which the evaluator could not clearly classify due to the language used in the site, the subject matter, or the fact that a Blog or Directory can not fall simply into one of the U/T categories. (Not every blog contains the string "blog" in their URL.)

Our experiments showed that the quality of the starting site was a very good predictor for the quality of the BCC sites. There were almost no trustworthy sites in the trust graph of sites U-1 to U-6. As one might expect, a trustworthy site is unlikely to deliberately link to an untrustworthy site, or even to a site that "associates" itself with an untrustworthy one. In other words, the "vote of confidence" analogy holds true for sites that are responsibly choosing their links.

Not surprisingly, the statement is not as strong when starting from a trustworthy site, since untrustworthy sites are free to link to whomever they choose. After all, there is some value in portraying a site in good company. Yet, spammers are unlikely to want to link to too many sites outside their spamming network in order to avoid "leaking" PageRank [5].

Research in the past has focused on the identification of web communities through the use of bipartite cores [29] or maximum flow in dense subgraphs [14]. These ideas do not apply to our construction. For one, we are not trying to identify a community of the starting site, but a sample of its trust neighborhood. In fact, we never look at the links coming out of s or any other site. One of the benefits of our method is that we do not need to explore the web graph explicitly, which would be impossible for a client computer.

7. RELATED WORK

Web spamming has received a lot of attention lately [1, 3, 4, 5, 12, 13, 20, 22, 23, 25, 29, 32, 33, 35]. The first papers to raise the issue were [33, 23]. The spammers' success was noted in [4, 10, 12, 13, 17, 24]. Web search was explained in [2]. The related topic of cognitive hacking was introduced in [11].

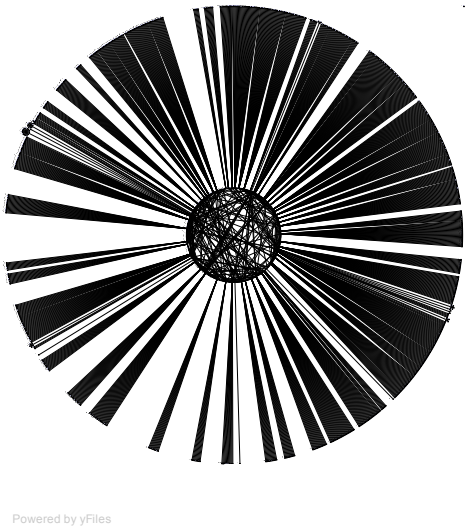

Figure 3: The trust graph of starting site U-1. The circularly drawn nodes in the middle form its largest biconnected component. This experiment found a trust graph of 1307 sites, 228 of which were connected with 465 edges into a BCC. No trustworthy sites were found in the BCC, while an estimated 65% of them were untrustworthy.

S	$\|V_G\|$	$\|E_G\|$	$\|V_{BCC}\|$	$\|E_{BCC}\|$	**Trust.**	**Untr.**
U-1	1307	1544	228	465	0%	65%
U-2	1380	1716	266	593	0%	88%
U-3	875	985	97	189	0%	95%
U-4	457	509	63	115	0%	83%
U-5	716	807	105	189	0%	73%
U-6	312	850	228	763	9%	57%
T-1	1429	1566	164	273	75%	1%
T-2	241	247	13	17	80%	13%

Table 2: Sizes of the explored graphs and their BCC's for six untrustworthy (U-1 to U-6) and two trustworthy (T-1 and T-2) starting sites. Column $|V_G|$ contains the number of vertices that our algorithm found in the trust neighborhood of starting site s (starting from site s and exploring in breadth-first search the backlinks of s. Column $|E_G|$ has the number of edges in the trust neighborhood. Columns $|V_{BCC}|$ and $|E_{BCC}|$ contains the numbers of edges of the largest biconnected component within G. The last two columns contain the estimated percentages of trustworthy and untrustworthy sites found in the BCCs. 20% of each BCC were sampled.

Characteristics of spamming sites based on diversion from power laws are presented in [12]. Current tricks employed by spammers are detailed in [19]. An analysis of the popular PageRank method employed by many search engines today and ways to maximize it in a spamming network is described in [5]. TrustRank, a modification to the PageRank to take into account the evaluations of a few seed pages by human editors, employees of a search engine, is presented in [20].

A comprehensive treatment on social networks is presented in [40]. The connection between the Web and social networks was explicitly noted in [30, 36] and implicitly used in [6, 28]. In fact, Kleinberg's work explores many of these connections (e.g., [27]). Identification of web communities was explored in [29, 14]. Propagation methods for trust and distrust are discussed in [18]. Work on topic-sensitive and personalized web search is presented in [21, 26]. The effect that search engines have on page popularity was discussed in [9].

8. CONCLUSIONS

In this paper we have argued that web spam is to cyberworld what propaganda is to society. As far as we know, this is the first time this relationship is noted. As evidence of the importance of this analogy, we have shown that the evolution of search engines can be simply understood as the search engines' response defending against spam. New search engines are not invented every few years, as it is sometimes reported; they are developed when researchers have a good answer to spam.

Further, our findings suggests that anti-spamming techniques can now be developed by mimicking anti-propagandistic methods. In particular, we have presented automatic ways of recognizing trust graphs on the web based on the biconnected component around some starting site. Experimental results from a number of such instances show our algorithm's ability of recognizing parts of a spamming network.

With such results, the question arises as to what one should do once one recognizes a spamming network. This is a question that has not attracted much attention in the past. The default approach is that a search engine would delete such networks from its indices [12] or might downgrade them by some prespecified amount [20].

Both of these approaches, however, require a universal agreement of what constitutes spam. Such an agreement cannot exist; one person's spam may be another person's treasure. Should the search engines determine what is trustworthy and what is not? Willing or not, they are the *de facto* arbiters of what information users see. As in a popular cartoon, a kid responds to the old man who has been looking all his life for the meaning of life: "If it is not on Google or eBay, it does not exist."

We believe that it is the users' right and responsibility to decide what is acceptable for them. Their browser, their window to cyberworld, should enhance their ability to make this decision. User education is fundamental: People should know how search engines work and why, and how information appears on the web. But they should also have a browser that can help them determine the validity and trustworthiness of information.

The tool we described in an earlier section is a first step in this direction. Ultimately, it would be used along with a set of trust certificates that contains the portable trust preferences of the user, a set of preferences that the user

can accumulate over time. A combination of search engines capable of providing indexed content and structure [21], including identified neighborhoods, with a browser capable of filtering those neighborhoods through the user's trust preferences, would provide a new level of reliability to the user's information gathering. Sharing ranking decisions with the end user will make it much harder for spammers to tune to a single metric.

8.1 Future Work

In our experiments we also devised a simple method to evaluate the similarity of the contents of each site to the starting site s. After the trust neighborhood was explored, we fetched and concatenated a few pages from each site (randomly choosing from the links that appeared in the top URL) into a document. Then, we tried to determine the similarity of each such document to the document of the starting site. Similarity was determined using the $tf.idf$ ranking on the universe of the sites explored. We are aware that having a limited universe of documents does not give the best similarity results, but we wanted to get a feeling of whether our method could further be used to distinguish between "link farms" and "mutual admiration societies". Though the initial results were encouraging (see Fig. 4), more work is needed in this area.

Several possible extensions can be considered in this work. Generating graphs with more backlinks per site, comparing the evolution of trust neighborhoods over time, examining the density of the BCCs, and finding a more reliable way to compute similarity are some of them. We also expect that the results would be strengthened if one considers the triconnected (or higher) components of the trust neighborhood.

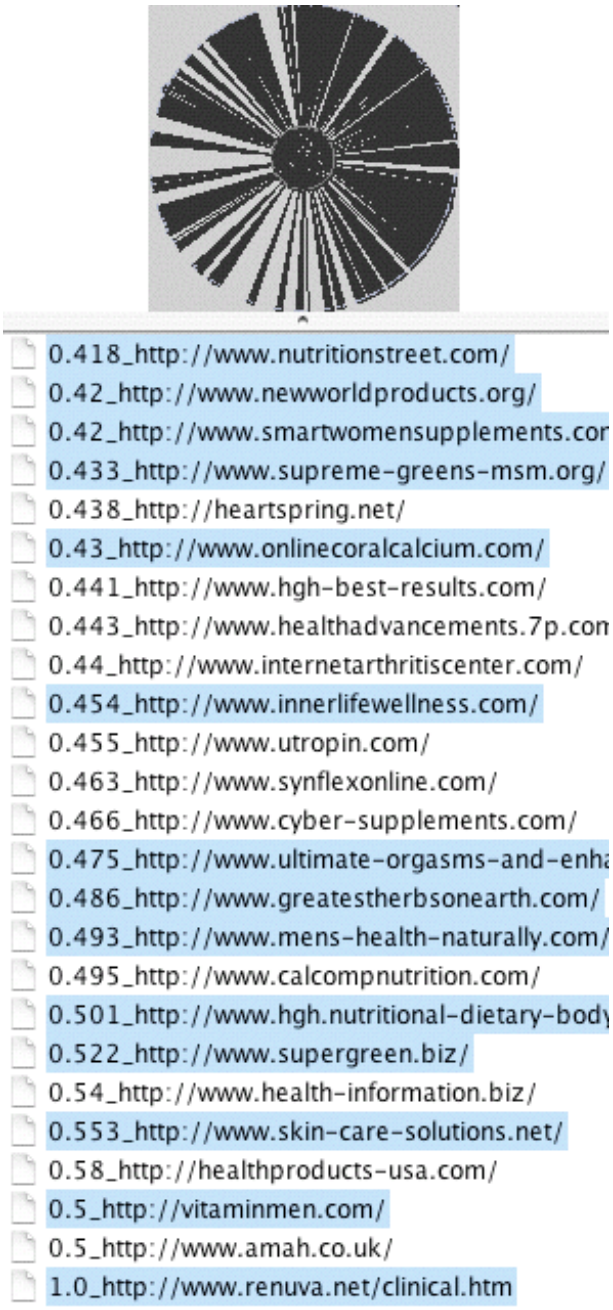

Figure 4: The list of sites similar to the starting site U-1. The hilited sites are those that participate in the BCC. The number in front of the URL corresponds to its calculated similarity to the starting site.

9. ACKNOWLEDGEMENTS

The authors would like to thank Mirena Chausheva and Meredith Beaton-Lacoste and Scott Dynes for their valuable contributions. They would also like to thank David "Pablo" Cohn and the anonymous referees of an earlier version of the paper for their suggestions. The graphs shown in this paper were drawn using the yEd package [41].

10. REFERENCES

[1] B. Amento, L. Terveen, and W. Hill. Does authority mean quality? Predicting expert quality ratings of web documents. In *Proceedings of the Twenty-Third Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2000.

[2] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the web. *ACM Transactions on Internet Technology*, 1(1):2–43, June 2001.

[3] K. Bharat, A. Z. Broder, J. Dean, and M. R. Henzinger. A comparison of techniques to find mirrored hosts on the WWW. *Journal of the American Society of Information Science*, 51(12):1114–1122, 2000.

[4] K. Bharat, B.-W. Chang, M. R. Henzinger, and M. Ruhl. Who links to whom: Mining linkage between web sites. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 51–58. IEEE Computer Society, 2001.

[5] M. Bianchini, M. Gori, and F. Scarselli. PageRank and web communities. In *Web Intelligence Conference 2003*, Oct. 2003.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998.

[7] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[8] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Comput. Networks*, 33(1-6):309–320, 2000.

[9] J. Cho and S. Roy. Impact of search engines on page popularity. In *WWW 2004*, May 2004.

[10] T. S. Corey. Catching on-line traders in a web of lies: The perils of internet stock fraud. Ford Marrin Esposito, Witmeyer & Glesser, LLP, May 2001. http://www.fmew.com/archive/lies/.

[11] G. Cybenko, A. Giani, and P. Thompson. Cognitive hacking: A battle for the mind. *Computer*, 35(8):50–56, 2002.

[12] D. Fetterly, M. Manasse, and M. Najork. Spam, damn spam, and statistics. In *WebDB2004*, June 2004.

[13] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the twelfth international conference on World Wide Web*, pages 669–678. ACM Press, 2003.

[14] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization of the web and identification of communities. *IEEE Computer*, 35(3):66–71, 2002.

[15] I. for Propaganda Analysis. How to detect propaganda. *Propaganda Analysis*, 1(2), 1937.

[16] Google. The Google API. http://www.google.com/apis/.

[17] L. Graham and P. T. Metaxas. "Of course it's true; i saw it on the internet!": Critical thinking in the internet era. *Commun. ACM*, 46(5):70–75, 2003.

[18] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *WWW 2004*, May 2004.

[19] Z. Gyongui and H. Garcia-Molina. Web spam taxonomy. Technical Report TR 2004-25, Stanford University, 2004.

[20] Z. Gyongui, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *VLDB 2004*, Aug. 2004.

[21] T. H. Haveliwala. Topic-sensitive pagerank. In *Proceedings of the eleventh international conference on World Wide Web*, pages 517–526. ACM Press, 2002.

[22] M. R. Henzinger. Hyperlink analysis for the web. *IEEE Internet Computing*, 5(1):45–50, 2001.

[23] M. R. Henzinger, R. Motwani, and C. Silverstein. Challenges in web search engines. *SIGIR Forum*, 36(2):11–22, 2002.

[24] M. Hindman, K. Tsioutsiouliklis, and J. Johnson. Googlearchy: How a few heavily-linked sites dominate politics on the web. In *Annual Meeting of the Midwest Political Science Association*, April 3-6 2003.

[25] L. Introna and H. Nissenbaum. Defining the web: The politics of search engines. *Computer*, 33(1):54–62, 2000.

[26] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the twelfth international conference on World Wide Web*, pages 271–279. ACM Press, 2003.

[27] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170. ACM Press, 2000.

[28] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632, 1999.

[29] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the Web for emerging cyber-communities. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1481–1493, 1999.

[30] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The web and social networks. *IEEE Computer*, 35(11):32–36, 2002.

[31] A. M. Lee and E. B. Lee(eds.). *The Fine Art of Propaganda*. The Institute for Propaganda Analysis. Harcourt, Brace and Co., 1939.

[32] C. A. Lynch. When documents deceive: trust and provenance as new factors for information retrieval in a tangled web. *J. Am. Soc. Inf. Sci. Technol.*, 52(1):12–17, 2001.

[33] M. Marchiori. The quest for correct information on the web: hyper search engines. *Comput. Netw. ISDN Syst.*, 29(8-13):1225–1235, 1997.

[34] M. L. Maulding. Lycos: Design choices in an internet search service. *IEEE Expert*, January-February(12):8–11, 1997.

[35] G. Pringle, L. Allison, and D. L. Dowe. What is a tall poppy among web pages? In *Proceedings of the seventh international conference on World Wide Web 7*, pages 369–377. Elsevier Science Publishers B. V., 1998.

[36] P. Raghavan. Social networks: From the web to the enterprise. *IEEE Internet Computing*, 6(1):91–94, 2002.

[37] G. Salton. Dynamic document processing. *Commun. ACM*, 15(7):658–668, 1972.

[38] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[39] A. Vedder. Medical data, new information technologies and the need for normative principles other than privacy rules. In *Law and Medicine. M. Freeman and A. Lewis (Eds.), (Series Current Legal Issues)*, pages 441–459. Oxford University Press, 2000.

[40] S. Wasserman and K. Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[41] yWorks. yEd – java graph editor, v. 2.2.1. http://www.yworks.com/en/products_yed_about.htm.